

Classification and keyword extraction of online harassment text in Thai social network

Siranuch Hemtanon¹, Ketsara Phetkrachang², Wachira Yangyuen³

¹Program of Management, Faculty of Business Administration, Rajamangala University of Technology Srivijaya, Songkha, Thailand

²Computer Engineering, Faculty of Engineering, Rajamangala University of Technology Srivijaya, Songkla, Thailand

³College of Industrial Technology and Management, Rajamangala University of Technology Srivijaya, Nakhon Si Thammarat, Thailand

Article Info

Article history:

Received Feb 5, 2023

Revised May 4, 2023

Accepted May 24, 2023

Keywords:

Classification

Cyberbullying

Keyword detection

Online harassment post

Social network post

Text mining

ABSTRACT

Online harassment in social network services (SNS) is a type of cyberbullying issue that needs to be addressed and required preventive measures. In this paper, we develop a detection of cyberbullying regarding harassment textual posts in Thai on the Facebook SNS. We collect public posts and ask experts to label the post as positive or negative regarding harassment posts or not. The annotated data are trained for binary classification considering words in the centre as features to predict malicious intent to insult and threaten other users. The information gain score obtained in generating a prediction model is ranked for the top 20 words with the highest score as significant words involving online harassment. From experiments, the results show that the detection performance obtained a 0.78 f1 score on average. The result analysis indicated that the word surface approach helps detect insulting post decently, but some posts with metaphor to tone down the malicious intent may not be detected as harmful semantic intent are hidden behind word form. Top-20 significant words for bullying showed that bullying posts were body-shaming and lower social status.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Ketsara Phetkrachang

Computer Engineering, Faculty of Engineering, Rajamangala University of Technology Srivijaya

Ratchadamnoennok Road, Bo Yang Subdistrict, Mueng District, Songkla, Thailand

Email: Ketsara.p@rmutsv.ac.th

1. INTRODUCTION

Bullying is the aggressive behavior of a person or a group toward another person over time. It results in severe and long-term effects, especially on mental health. Since online communication tools have become a ubiquitous and essential part of our lives, some utilize devices and applications for acting maliciously, or menacing toward others in the form of cyberbullying [1], [2]. As convenient from online software, cyberbullying can happen at any time in public or in private as only known to the target and the person bullying [3]–[5]. One of the standard cyberbullying methods is to post a text meant to insult, embarrass, or threaten a victim on a social network.

As cyberbullying occurs globally and impacts the mental health of victims, detection of cyberbullying is essential as a preventive measure, especially on the social network. A cyberbully detection system [6] has become a topic for research aiming to identify and classify online activities meant to harass other social network users, such as flaming, insulting, and terrorizing. The standard method is to apply text mining techniques [7], [8] or automated classification [9]–[11] to categorize words with malicious intent on infamous social networks platforms such as Twitter and Reddit.

Despite existing works of cyberbullying detection on English messages in social networks, it may not directly apply to other languages with a different culture. The meaning of insulting and embarrassing differs

from culture to culture based on what people of the culture hold importance. In this paper, we aim to detect online Thai textual harassment and extract the keywords of online harassment. This work's target social network service (SNS) is Facebook, the most used SNS in Thailand. The rest of this paper is organized as follows. Section 2 provides background knowledge about cyberbullying and existing research works related to this work. Section 3 describes the details of online harassment detection of Thai text in Thai on social network platforms. Section 4 gives the experiment setting and evaluation results. Lastly, section 5 provides a conclusion and future work.

2. BACKGROUND

2.1. Cyberbullying

Cyberbullying is conceptualized as the intentional use of electronic resources, including mobile phones, computers, and other electronic communication devices via online SNSs, to taunt, hurt, threaten, embarrass, and harm others [12], [13]. Cyberbullying can be an extension of concurrent physical bullying or a separate incident, specifically on online SNS. The problem of cyberbullying is spreading globally as more people have access to SNS [14] and has become a concern as a cause of mental issues [15]–[17], such as depression syndrome [18], and suicidal incidents. There are several actions that can be counted as cyberbullying, and those can be grouped as follows:

- Exclusion; is a group action of deliberately leaving someone out to make them feel excluded. It exists in both real-life and online settings. Most cases are a continuation of real-life bullying situations.
- Framing; is an act of posting making up a story or false information to their target. This action is an online version of spreading accusive rumors to discredit a target.
- Faking; is creating an account pretending to be a target and posting inappropriate things to frame a target. This online action may affect the target's reputation and directly harm one's social status.
- Harassment; is a category generally referring to a use of hurtful or threatening online post and messages with the intention of doing harm towards a target.

Among the above actions, some are actions not limited to the online SNS but a continuation of normal bullyings, such as exclusion and framing. Some actions like exclusion cannot be easily detected from the outside as they are only known within an internal group. However, one of the cyberbullying occurrences that can be noticed in general is online harassment which is an act of making a public post to embarrass, insult, and threaten a target. As cyberbullying is harmful and continuously increases, preventive measures and healing of the affected targets are recommended.

2.2. Related work on text mining

Cyberbullying detection towards the prevention of cyberbullying, several works proposed using a text-mining approach to detect online posts containing bullying words. Noviantho *et al.* [19] developed a system for cyberbullying classification using text mining. They selected naive bayes and support vector machine (SVM) for classification on n-grams of 1 to 5 words. The selected dataset is a cyberbullying conversation in English. They achieved an accuracy score of over 92%. Nalini and Sheela [20] proposed classification using a text classifier to detect cyberbullying tweets. In their work, English words in tweets (posts on Twitter) are considered features for text analysis for classification. They applied a weighting scheme as feature selection to help detect features with malicious intent. Dadvar *et al.* [21] proposed a cyberbullying detection based on English datasets from MySpace. Instead of using text data, this work used the gender feature provided in a user profile to enhance the discrimination capacity of a classification. Later, Dadvar and Heidari [22] improved their work by including useful information such as age and gender as features for classification. With additional information, classification performance was improved in a tradeoff that the method can only be applied to online users who provide such information publicly.

In 2021, Thai-textual cyberbullying detection [23] was proposed. The primary classification technique was SVM with term frequency for feature selections. The applied dataset collected bullying posts against celebrities from several social network sources, including Facebook, Instagram, Twitter, and YouTube. With the comments towards celebrities, the insulting words thus mostly were directed towards their behavior in general.

3. THAI ONLINE TEXT HARASSMENT DETECTION

This paper consists of two tasks which are a text classification and keyword extraction. An overview is as shown in Figure 1. The training data in this work is pre-processed Thai text with cyberbullying annotation. The classification model is then generated for detecting a textual post intended for harassment. The model then is analyzed for words having important high scores as bullying keywords.

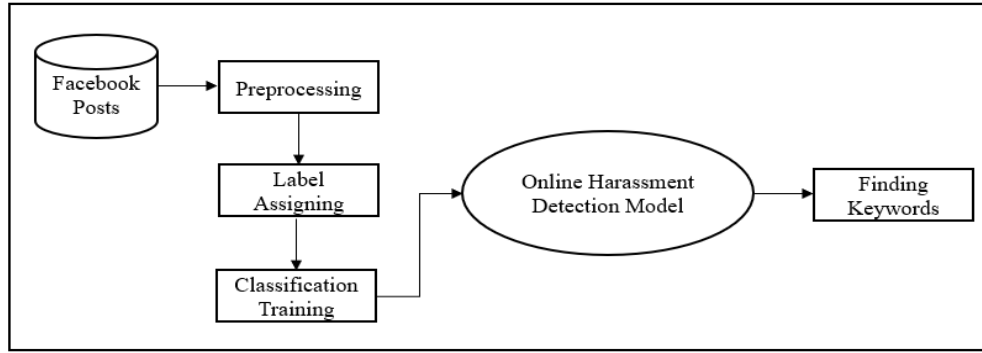


Figure 1. An overview of the classification and keyword finding of online harassment text in Thai on SNS

3.1. Text data preparation

In this work, the data are Thai text data collected from public posts on Facebook. Only texts are recorded without name or identification under the research license id WUEC-23-039-01. The collected texts are to be processed as follows. The typos and misspellings words are corrected to reduce noises, and scattering of the same word. Furthermore, word segmentation is applied to define terms in the given text. The word segmentation service selected in this work is Lexto-plus [24] for the longest word. However, the word segmentation performance can be incorrect from ambiguity and unknown words, especially the new word that recently emerged among teenagers. Thus, post-edit is applied to correct word segmentation for maintaining input quality. Last, functional words representing grammatical functions with little to no meaning are removed to maximize text processing performance in terms of computational complexity from lowering search space.

The training data in this work is pre-processed Thai text with cyberbullying annotation. The classification model is then generated for detecting a textual post intended for harassment. The model then is analyzed for words having important high scores as bullying keywords.

Once the preparation is done, the terms that appear in Facebook posts are formed into a vector representation. The terms and posts are aligned into a vector regarding their existence. The word vector will then be used in a later learning process to create a classification model. To assign a class to each textual post, we ask healthcare personnel and linguist to annotate if the text is bullying or not, regarding insulting and threatening as 'positive' and 'negative' for containing bullying content and not containing, respectively. The annotation, thus, is a binary class for classification.

3.2. Classification model

The model for classification in this work is based on a supervised learning technique. With the labelled data, the task is binary classification, and several machine-learning techniques can be applied to generate a classification model [25]. In this work, we select the decision tree (DT) ID3 technique [26] as our method to generate a classification model. Words in a vector are used as a list of features to determine a class of being online harassment or not. One benefit of DT is that it allows us to see how the decision is made explicitly, and the tree can be further analyzed.

3.3. Keyword of being cyberbullying

In generating a classification model of DT, information gain is calculated to measure how much information a feature provides about a class. In this work, features are words that appear in a training text. Hence, words can be ranked following their obtained IG score to represent how much impact the words signify the positive of cyberbullying. IG score is calculated to measure the difference in entropy values from before to after partitioning the set regarding a word A by (1):

$$IG(A) = H(S) - \sum_{t \in T} p(t)H(t) = H(S) - H(S|A) \quad (1)$$

where $H(S)$ is entropy of set S ; T refers to subsets from separating set S by a feature A , and $p(t)$ is a proportion of the number of items in subset t to the number of items in set S . Last, $H(S)$ refers to entropy of the subset t . The higher the IG score of a word, the more significant the word leads to be positive for cyberbullying. In this work, we use the IG score to rank the top of words as a list of keywords for leading to incurring online harassment intention regarding cyberbullying.

4. RESULTS AND DISCUSSION

4.1. Data collection

The target textual data in this experiment are Thai texts publicly posted on Facebook SNS. We randomly collected 12,000 posts made between April 2022 to August 2022. We then asked healthcare personnel and a linguist to assign the label to the posts. The criteria for keeping the post for representing cyberbullying regarding harassment and terrorizing are as follows. For the posts that both experts agreed on having harassment intent and not having harassment intent, we keep them for the ‘positive’ and ‘negative’ classes, respectively. The posts that were disagreed by both experts were discarded.

Furthermore, the posts containing less than 5 words and longer than 40 words were also discarded since the content was too insufficient regarding the semantics of the content or contained multiple concepts to determine clear intention. As a result, there were 3,440 posts containing the intent of cyberbullying and 4,219 general posts with ‘positive’ and ‘negative’ labels, respectively. For data statistics, there were 8,158 unique words from the remaining posts. The average post length was 14.89 words.

4.2. Evaluation results of detecting online harassment posts

The collected data were prepared for 5-fold cross-validation for evaluation. Each separated fold will surely contain the same number of positive and negative depression instances. The evaluation measurements in both experiments are precision (P), recall (R), and F-measure (F1) scores. The results of detecting online harassment posts of each fold are given in Table 1.

Table 1. Evaluation results of detecting online harassment

	P	R	F1
Fold-1	0.75	0.81	0.78
Fold-2	0.77	0.80	0.78
Fold-3	0.82	0.83	0.82
Fold-4	0.75	0.78	0.76
Fold-5	0.74	0.77	0.75
Average	0.77	0.80	0.78

The results were not impressively high as there were some incorrect predictions. From analysis, we found that most of the correct results were in the type of insulting posts and posts for embarrassing a target. On the other hand, threatening posts gave the most incorrect predictions since the words in such posts were common and not specified enough to distinguish them from normal posts regarding terrible news. Furthermore, we also noticed that the threatening training posts often used metaphoric style to prevent direct semantics of malicious intent or to tone down the meaning from the public, such as “do you want to smell dirt” to refer to getting beaten to the ground where one can smell the dirt directly. Thus, these posts are manageable with the currently applied technique, which uses word surface to determine to bully.

4.3. Found keywords regarding online harassment

To rate which words are significant in deciding on having online harassment intention or not, the IG score in generating the DT model is considered. We trained a model with all text posts in the dataset without fold separation for this case. The top 20 words with the highest IG score are listed in Table 2 in descending order. The words are in Thai, so we give part of speech (POS) and literal translation to help with understandability. This word list contains terms containing meaning intending to bully others. Most of them are adjectives to describe the negative meaning, especially body shaming (rank#3, 7, 9, 17, and 20) and social-status shaming (rank#4, 9, 11, and 19).

Table 2. The top-20 words with highest IG score

Rank	Word (POS) [Lit. trans]	Rank	Word (POS) [Lit. trans]
1	ดอแหล (v and adj)	11	หวั่งสูง (v)
2	สฤน (adj and adv)	12	นำหนัก (n)
3	ปลอม (adj)	13	เยอะ (adj and adv)
4	ตลาดล่าง (n and adj)	14	เหม็นเปรี้ยว (adj)
5	สันดาน (n and adj)	15	ขี้เวร (adj and adv)
6	อีดอก (n)	16	วอก (adj)
7	จิว (n and adj)	17	หนาลอย (adj)
8	ฮึ้น (pron)	18	สะดอ (v)
9	แบกปูน (v and adj)	19	สะเออะ (v)
10	เกินเบอร์ (adj and adv)	20	หน้าปลวก (adj)

4.4. Discussion

From the experiment, the most common harsh online posts were body-shaming and insulting social status. These signify that bullying people may target body and social status as their main harassment content, and most of the targets are those with poor status and non-traditional body shapes. These harsh words are not only used in online SNS but are also used in real-life conversations. Thus, it is arguable that they bring the same bullying attitudes online. The experiment results also showed that using word surface may not be able to fully detect online harassment since Thai culture tends to tone down the intention of posts by using metaphors. However, this can be understandable for a person using real-world knowledge, but such metaphor posts remain a challenge in natural language processing for Thai.

As cyberbullying may include more types, such as excursion, framing, and faking, detecting those activities online is also essential for online bullying prevention. Since these actions are not about wording alone, a word-based text mining approach is insufficient for detecting bullies online. It requires extra information to properly distinguish the activities, such as behavior patterns (user behavior within SNS) for the excursion and user profile information and post history for faking and framing.

5. CONCLUSION AND FUTURE WORK

In this paper, we develop a detection of cyberbullying regarding harassment textual posts in Thai on the Facebook SNS. By collecting public posts and giving them the label, we train a classification model based on words in the post as features to predict malicious intent to insult and threaten other users. In generating a classification model, the information gain score is calculated for words that may signify bullying intention, and we list the top 20 words with the highest score. The results show that the detection performance obtained a 0.78 f1 score. The analysis indicated that the word surface approach might help detect insulting but not threatening posts with metaphors to tone down the malicious intent. Top-20 significant words for bullying showed that bullying posts were body-shaming and lower social status. To improve our work, we plan to include behavior patterns for detecting excursion-type cyberbullying, user profile information, and post history for faking and framing on the online platform. We also plan to research metaphoric semantics to cover Thai-style metaphors in detecting insulting and threatening online posts.




REFERENCES

- [1] C. E. Sanders, "What is Bullying?," in *Bullying*, Elsevier, 2004, pp. 1–16, doi: 10.1016/B978-012617955-2/50004-7.
- [2] S. Einarsen and M. B. Nielsen, "Workplace bullying as an antecedent of mental health problems: a five-year prospective and representative study," *Int Arch Occup Environ Health*, vol. 88, no. 2, pp. 131–142, Feb. 2015, doi: 10.1007/s00420-014-0944-7.
- [3] N. E. Willard, *Cyberbullying and cyberthreats: Responding to the challenge of online social aggression, threats, and distress*. in Cyberbullying and cyberthreats: Responding to the challenge of online social aggression, threats, and distress. Champaign, IL, US: Research Press, 2007, pp. v, 311.
- [4] J. Snakenborg, R. Van Acker, and R. A. Gable, "Cyberbullying: Prevention and Intervention to Protect Our Children and Youth," *Preventing School Failure: Alternative Education for Children and Youth*, vol. 55, no. 2, pp. 88–95, Jan. 2011, doi: 10.1080/1045988X.2011.539454.
- [5] K. L. Mason, "Cyberbullying: A preliminary assessment for school personnel," *Psychol. Schs.*, vol. 45, no. 4, pp. 323–348, Apr. 2008, doi: 10.1002/pits.20301.
- [6] A. Ademiluyi, C. Li, and A. Park, "Implications and Preventions of Cyberbullying and Social Exclusion in Social Media: Systematic Review," *JMIR Form Res*, vol. 6, no. 1, p. e30286, Jan. 2022, doi: 10.2196/30286.
- [7] S. M. Inzalkar and J. Sharma, "A survey on text mining- techniques and application," *International journal of research in science and engineering*, no. Techno-Xtreme 16, pp. 488–495, 2015.
- [8] R. Agrawal and M. Batra, "A detailed study on text mining techniques," *International Journal of Soft Computing and Engineering (IJSCE)*, vol. 2, no. 6, pp. 118–121, Jan. 2013.
- [9] C. H. Caldas, L. Soibelman, and J. Han, "Automated Classification of Construction Project Documents," *J. Comput. Civ. Eng.*, vol. 16, no. 4, pp. 234–243, Oct. 2002, doi: 10.1061/(ASCE)0887-3801(2002)16:4(234).
- [10] C. H. Brown, E. W. Holman, S. Wichmann, and V. Velupillai, "Automated classification of the world's languages: a description of the method and preliminary results," *Language Typology and Universals*, vol. 61, no. 4, pp. 285–308, Nov. 2008, doi: 10.1524/stuf.2008.0026.
- [11] J. C.-Huang, R. Settini, X. Zou, and P. Solc, "Automated classification of non-functional requirements," *Requirements Eng*, vol. 12, no. 2, pp. 103–120, Apr. 2007, doi: 10.1007/s00766-007-0045-1.
- [12] I.-K. Peter and F. Petermann, "Cyberbullying: A concept analysis of defining attributes and additional influencing factors," *Computers in Human Behavior*, vol. 86, pp. 350–366, Sep. 2018, doi: 10.1016/j.chb.2018.05.013.
- [13] P. K. Smith, J. Mahdavi, M. Carvalho, S. Fisher, S. Russell, and N. Tippett, "Cyberbullying: its nature and impact in secondary school pupils," *J Child Psychol and Psychiat*, vol. 49, no. 4, pp. 376–385, Apr. 2008, doi: 10.1111/j.1469-7610.2007.01846.x.
- [14] E. O. Okoiye, N. N. Anayochi, and T. A. Onah, "Moderating Effect of Cyber Bullying on the Psychological Well-Being of In-School Adolescents in Benin Edo State Nigeria," *EJSD*, vol. 4, no. 1, pp. 109–118, Feb. 2015, doi: 10.14207/ejsd.2015.v4n1p109.
- [15] Y. Akbulut and B. Eristi, "Cyberbullying and victimisation among Turkish university students," *AJET*, vol. 27, no. 7, pp. 1155–1170, Nov. 2011, doi: 10.14742/ajet.910.
- [16] R. Chisholm M., *Person and Object: A Metaphysical Study*, 1st Edition. Routledge Taylor and Francis Group, 2014.
- [17] J. Rivituso, "Cyberbullying Victimization among College Students: An Interpretive Phenomenological Analysis," *Journal of Information Systems Education*, vol. 25, no. 1, pp. 71–76, Jan. 2014.




- [18] I. Pantic, "Online Social Networking and Mental Health," *Cyberpsychology, Behavior, and Social Networking*, vol. 17, no. 10, pp. 652–657, Oct. 2014, doi: 10.1089/cyber.2014.0070.
- [19] Noviantho, S. M. Isa, and L. Ashianti, "Cyberbullying classification using text mining," in *2017 1st International Conference on Informatics and Computational Sciences (ICICoS)*, Semarang: IEEE-2017 1st International Conference on Informatics and Computational Sciences (ICICoS), Nov. 2017, pp. 241–246, doi: 10.1109/ICICOS.2017.8276369.
- [20] K. Nalini and L. J. Sheela, "Classification of Tweets Using Text Classifier to Detect Cyber Bullying," *Proceedings of the 49th Annual Convention of the Computer Society of India CSI*, Springer, Cham, 2015, vol. 2, pp. 637–645, doi: 10.1007/978-3-319-13731-5_69.
- [21] M. Dadvar, F. M. G. de Jong, R. J. F. Ordelman, and R. B. Trieschnigg, "Improved cyberbullying detection using gender information," in *Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012)*, Ghent University, pp. 23–25, Feb. 2012.
- [22] E. Dadvar and A. Heidari, "A Review on Separation Techniques of Graphene Oxide (GO)/Base on Hybrid Polymer Membranes for Eradication of Dyes and Oil Compounds: Recent Progress in Graphene Oxide (GO)/Base on Polymer Membranes-Related Nanotechnologies," *Clin Med Rev Case Rep*, vol. 5, no. 8, pp. 228–247, Aug. 2018, doi: 10.23937/2378-3656/1410228.
- [23] L. Mookdarsanit and P. Mookdarsanit, "ThaiWritableGAN: Handwriting Generation under Given Information," *IJCDS*, vol. 10, no. 1, pp. 689–699, May 2021, doi: 10.12785/ijcds/100165.
- [24] C. Haruechaiyasak and A. Kongthon, "LexToPlus: A Thai Lexeme Tokenization and Normalization Tool," *Nagoya, Japan: The 4th Workshop on South and Southeast Asian NLP (WSSANLP), International Joint Conference on Natural Language Processing*, pp. 14–18, Oct. 2013.
- [25] S. Hemtanon, S. Aekwarangkoon, and N. Kittiphattanabawon, "Detection of Depression-Positive Thai Facebook Users Using Posts and Their Usage Behavior," *International Conference on Computing and Information Technology*, Springer, Cham, 2021, vol. 251, pp. 77–87, doi: 10.1007/978-3-030-79757-7_8.
- [26] P. H. Swain and H. Hauska, "The decision tree classifier: Design and potential," *IEEE Trans. Geosci. Electron.*, vol. 15, no. 3, pp. 142–147, Jul. 1977, doi: 10.1109/TGE.1977.6498972.

BIOGRAPHIES OF AUTHORS






Siranuch Hemtanon    holds a Ph.D. in Management Information Technology from Walailuk University, Thailand. She is currently instructor of Management Program, Faculty of Business Administration, Rajamangala University of Technology Srivijaya, no. 1 Ratchadamnoen Nok Road, Bo Yang Subdistrict, Mueang District, Songkhla 90000, Thailand. Her research interests include data mining, natural language processing, machine learning, and question answering systems. She can be contacted at email: Siranuch.h@rmuts.ac.th or Mitnuchie@gmail.com.



Ketsara Phetkrachang    received a Ph.D. in Management Information Technology from Walailuk University, Thailand, 2018. She is currently instructor of Computer Engineering Program, Faculty of Engineering, Rajamangala University of Technology Srivijaya in Songkla, Thailand no.1 Ratchadamnoen Nok Road, Bo Yang Subdistrict, Mueang District, Songkhla 90000, Thailand. Her research interests are information retrieval, information engineering, question answering systems, and expert system. She can be contacted at email: Ketsara.p@rmuts.ac.th or ketsara1782@gmail.com.



Wachira Yangyuen    holds a Ph.D. in Management Information Technology from Walailak University, Thailand. He is the instructor in Digital Entrepreneurship of College of Industrial Technology and Management, Rajamangala University of Technology Srivijaya. The college is located at 99 Moo4, Tongnien, Khanom, Nokhon Si Thammarat, South of Thailand. His research was interested in information behavior, data mining, and business information system. He can be contacted at email: wachira.y@rmuts.ac.th.